Characterization of a Large Sexually Dimorphic Genome Interval in *Salix purpurea* L. (Salicaceae)

Ran Zhou[*], David Macaya-Sanz[*], Eli Rodgers-Melnick[*], Craig H. Carlson[†], Fred E. Gouker[†], Luke M. Evans[*], Jeremy Schmutz[‡,**], Jerry W. Jenkins[‡], Juying Yan[**], Gerald A. Tuskan[**,††], Lawrence B. Smart[†], and Stephen P. DiFazio*

* Department of Biology, West Virginia University, Morgantown, WV 26506-6057

† Horticulture Section, School of Integrative Plant Science, Cornell University, New York State Agricultural Experiment Station, Geneva, NY 14456

‡ HudsonAlpha Institute of Biotechnology, 601 Genome Way Northwest, Huntsville, AL 35806 USA

** Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA

†† Biosciences Division, Oak Ridge National Lab

Running Title: Genomic Sex Dimorphism in *S. purpurea*

Key Words: sex, *Salix*, genome, suppressed recombination, dioecy

Corresponding Author:

Stephen P. DiFazio
Department of Biology
West Virginia University
53 Campus Drive
Morgantown, WV 26506-6057, USA
(304) 293-5314
spdifazio@mail.wvu.edu

Abstract

Dioecy has evolved numerous times in plants, but heteromorphic sex chromosomes are apparently rare. Sex determination has been studied in multiple *Salix* and *Populus* (Salicaceae) species, and *Populus* typically has an XY sex determination system on chromosome 19, while *S. suchowensis and S. viminalis* have a ZW system on chromosome 15. Here we use quantitative trait locus mapping and a genome-wide association study to demonstrate that *Salix purpurea* also has a ZW system on chromosome 15. This region is characterized by reduced recombination, high structural polymorphism, and an abundance of transposable elements, as expected for a sex chromosome. Genes from this region are known to be involved in sex expression in other plants. We also show that chromosome 19 has some sex chromosome characteristics in *S. purpurea*, including significant QTL and association peaks for sex. Due to a lack of such signatures on chromosome 15 in *Populus*, we hypothesize that sex determination was originally on chromosome 19 in this lineage.

Introduction

Nearly 90% of flowering plants are hermaphroditic (containing both male and female floral parts in the same flower), and less than 6% are dioecious (separate male and female individuals) (Renner 2014). Evolutionary factors favoring dioecy include inbreeding avoidance and the ability to maximize reproductive output through unisexual resource partitioning (Charlesworth and Charlesworth 1978; Charnov 1982; Ashman 2006). In angiosperms, dioecy has independently evolved hundreds of times from hermaphroditic progenitors (Renner 2014). Evolutionary pathways to dioecy include gynodioecious, heterostylous, and monoecious intermediates (Lloyd 1979; Ainsworth 2000; Charlesworth 2006), but monoecious intermediates tend to be the most common mechanism in woody angiosperms (Olson *et al.* 2017).

Trait divergence between females and males can be facilitated by the presence of sex chromosomes, as these are the only genomic regions that consistently differ between the sexes (Rice 1984; Mank 2009; Barrett and Hough 2013). Sex chromosomes usually have suppressed recombination and increased haplotype divergence due to independently accumulating mutations, leading to the development of sexually dimorphic regions (SDR, regions that consistently differ between males and females). The SDR may comprise a majority of the chromosome or only a small portion. Heterogametic SDRs may confer either maleness (XY

system), as in *Silene latifolia*, *Carica papaya*, *Phoenix dactylifera*, *Diospyros lotus*, and *Populus trichocarpa*; or femaleness (ZW system), as in *Fragaria chiloensis*, *Silene ottites*, and *Pistacia vera* (reviewed in Charlesworth 2016; Vyskot & Hobza 2015). Sex chromosomes also contain pseudoautosomal regions (PAR) where sex chromosomes recombine freely and may often show elevated recombination (Nicolas *et al.* 2005; Otto *et al.* 2011). Many plant sex chromosomes are homomorphic, exhibiting no strong morphological differences, suggesting that these chromosomes are at an early stage of development (Ming and Moore 2007).

The Salicaceae family is an excellent model system for exploring the ecological and evolutionary dimensions of dioecy and sexual selection in plants. Widely distributed across temperate, boreal, and arctic regions of the globe, these genera represent a diverse assemblage of catkin-bearing trees and shrubs (Karp *et al.* 2011). There are approximately 30 *Populus* species, most of which are trees that grow in the northern hemisphere (Slavov and Zhelev 2010). In contrast, there are approximately 500 *Salix* species, most of which are shrubs (Dickmann and Kuzovkina 2014). Nearly all species in *Salix* and *Populus* are dioecious, but none have obvious heteromorphic sex chromosomes (Peto 1938). *Salix* is primarily insect pollinated (Karrenberg *et al.* 2002), and produces complex volatiles and nectar rewards (Füssel *et al.* 2007). In contrast, *Populus* is almost exclusively wind-pollinated. Furthermore, both lineages share a well-preserved whole genome duplication (Tuskan *et al.* 2006; Hou *et al.* 2016) and both show an ongoing propensity toward polyploid formation (Mock *et al.* 2012; Serapiglia *et al.* 2015), thus facilitating exploration of the relationship between polyploidy and sex chromosome evolution (Ashman *et al.* 2013; Glick *et al.* 2016).

There has been considerable work on characterizing sex determination in *Populus* over the past decade. The SDR has been mapped to the proximal telomeric end of chromosome 19 in *P. deltoides* and *P. nigra* (Gaudet *et al.* 2008; Yin *et al.* 2008) and to a pericentromeric region of chromosome 19 in *P. tremuloides*, *P. tremula*, and *P. alba* (Pakull *et al.* 2009; Paolucci *et al.* 2010; Kersten *et al.* 2014). In both *P. deltoides* and *P. alba*, the SDR was mapped on a female genetic map but not on a male genetic map, possibly supporting female heterogamety (Yin *et al.* 2008; Paolucci *et al.* 2010). In *P. tremuloides* and *P. nigra*, the SDR was mapped on the male genetic map and not on the female genetic map, suggesting male heterogamety (Gaudet *et al.* 2008; Kersten *et al.* 2014). Recently, a genome-wide association study (GWAS) on 52 *P. trichocarpa* and 34 *P. balsamifera* found 650 SNPs significantly associated with sex. These sex-

associated markers were nearly fixed heterozygous in males and homozygous in females, which is consistent with an XY sex-determination system (Geraldes *et al.* 2015). However, the significant marker associations were not confined to chromosome 19 but were scattered throughout the genome, possibly due to problems with assembly of the structurally-complex SDR (Geraldes *et al.* 2015).

In contrast to *Populus*, the SDR has been mapped to chromosome 15 in *S. viminalis* and *S. suchowensis* (Temmel *et al.* 2007; Hou *et al.* 2015; Pucholt *et al.* 2015). Furthermore, there is a preponderance of female heterozygosity in the SDR of these species, indicating a ZW sex determination system, in contrast to *Populus* (Hou *et al.* 2015; Pucholt *et al.* 2015). However, neither study identified candidate genes in the *Salix* SDR that were orthologous to genes in the SDR of *Populus* (Hou *et al.* 2015; Pucholt *et al.* 2015). Thus, *Salix* and *Populus* appear to have different sex determination mechanisms or sex-determining genes, and the nature of the SDR in the *Salicaceae* family remains poorly-characterized. In this study, we sought to explore the SDR in an additional Salicaceae species, *Salix purpurea*. Using robust linkage and association analyses, we show that the principal SDR is on chromosome 15, and that the genotype configuration in this region is consistent with a ZW system of sex determination. Furthermore, we also present evidence that chromosome 19 may retain a residual SDR that has been superseded by the chromosome 15 locus.

Materials and Methods

Genome Assembly

This work is based primarily on v1.0 of the *S. pupurea* genome, which is being described more completely in a separate publication (Smart et al., in preparation). Briefly, a female diploid genotype of *Salix purpurea* (clone 94006) was collected from the banks of the Fish Creek River in Upstate New York in 1994 (43.2168 N, -75.6333 W). This clone has been an important parent in *Salix* breeding programs, and is also the source of the reference genome that has been developed by the Joint Genome Institute and a consortium of researchers (available at http://phytozome.jgi.doe.gov). All DNA and RNA samples used for genomic and transcriptomic sequencing were derived from clonally propagated individuals of this genotype. ALLPATHS-LG was used to assemble sequences representing ~140X coverage of Illumina paired-end sequences,

as well as a set of mate-pair libraries (4.5 Kb, 5.3 Kb, 6.5 Kb), producing contigs with an L50=46 kb and scaffolds with L50=191 kb. The ALLPATHS-LG assembly has a total length of 348 Mb and a total span of 392 Mb (including gaps) but is still relatively fragmented due to a high level of heterozygosity (1 SNP per 120 bp, or 0.8%) and extensive structural variation. Assessment of the assembly quality against willow BACs and transcripts suggested that ~ 78% to 85% of the willow genome is captured in the current assembly. Gene annotations were accomplished using the Phytozome pipeline (Goodstein *et al.* 2012). The RepeatModeler (v1.0.8) package (http://www.repeatmasker.org) was used to identify and mask repetitive elements.

Genetic Mapping and Pseudomolecule Assembly

An F_1 mapping population was produced by crossing two *S. purpurea* accessions, clone 94006 (female) and clone 94001 (male), and intercrossing two of the resulting progeny (female 'Wolcott' and male 'Fish Creek') to produce over 500 F_2 progeny (referred to as Family 317). The parents and progeny, were genotyped via "Genotyping by Sequencing" (GBS) using *Eco*T221 and *Ape*KI restriction enzymes, and 96-fold multiplexed sequencing on an Illumina HiSeq Genome Analyzer (Elshire *et al.* 2011). SNPs were identified using the reference based pipeline of TASSEL (Glaubitz *et al.* 2014) using the *S. purpurea* v1.0 reference genome (available at http://phytozome.jgi.doe.gov). SNPs were also called using the UNEAK pipeline from TASSEL (Glaubitz *et al.* 2014). SNPs were filtered using the following parameters: -hetFreq 0.75 -mnTCov 0.01 -mnSCov 0.2 -mnMAF 0.05 -hLD -mnR2 0.2 -mnBonP 0.005, and <40% missing data. A total of 8,531 informative GBS markers obtained from 411 F_2 progeny were used to derive separate maps for markers in the three informative configurations: male backcross (n=2623), female backcross (n=2211), and intercross (n=3697). These genetic maps were integrated with the reference genome assembly to produce a combined map on which 276 Mb (70%) of sequence scaffolds were anchored, with intervening gaps that were proportional to distances between mapped markers in the scaffolds. The remaining unplaced scaffolds contained another 116 Mb of sequence. Assuming that the unplaced scaffolds are not alternative haplotypes of the mapped scaffolds, the total estimated genome size is approximately 392 Mb, an estimate that was corroborated by kmer counting and flow cytometry (Smart et al., in preparation). The assembly was compared to the *Populus trichocarpa* v3.0 reference genome with LASTZ (v1.03.66), using parameters to exclude alignments between paralogous segments derived from

the most recent shared whole genome duplication (gapped, chain, transition, maxwordcount=4, exact=100, step=20).

Identification of Sexually Dimorphic Genome Regions

Sex was scored for F_2 progeny by repeated observations during the spring of 2012, 2013, and 2015 in common gardens at the New York State Agricultural Experiment Station (Cornell University) in Geneva, NY. Quantitative Trait Locus (QTL) mapping was performed using the r/QTL package in R with a binary phenotype model (Arends *et al.* 2010). Logarithm of odds (LOD) support intervals or an approximate Bayesian credible interval were calculated using r/QTL. QTL mapping was performed for all three genetic maps (female backcross, male backcross and intercross).

We also performed a Genome-Wide Association Study (GWAS) on the sex trait using a population of unrelated individuals collected from the wild. A population of 112 *Salix purpurea* individuals was collected from upstate New York, Pennsylvania, Connecticut, and Vermont and planted in common gardens at Cornell University in Geneva, NY and at West Virginia University in Morgantown, WV. Sex was scored in the spring of 2013 and 2014 for six clonal replicates at each site. Only individuals for which sex was consistently and unambiguously scored as male or female were used for the analysis. The population was genotyped using GBS with the *Ape*KI restriction enzyme and 48-fold multiplex sequencing on an Illumina HiSeq Genome Analyzer. SNPs were called and filtered as described above, yielding 85,543 SNPs for analysis. A kinship matrix was calculated using the scaled Identity-by-State (IBS) method implemented in the EMMAX package (Kang *et al.* 2010). Clonal ramets were identified based on pairwise IBS values in comparison to pairwise IBS of the F_2 population described above (Figure S1). This resulted in removal of 34 ramets belonging to 9 clonal groups. Three apparently hermaphroditic individuals and 12 individuals with inconsistent sex phenotypes were also excluded from this analysis, leaving a total of 38 females and 22 males. To control for the influence of population structure, a Principal Components Analysis (PCA) was performed using smartPCA in the Eigenstrat package (Price *et al.* 2006). GWAS for sex was performed with the first two principal components and the kinship matrix as covariates using a mixed linear model implemented in the EMMAX package (Kang *et al.* 2010). We controlled for multiple testing using a Bonferroni correction with an alpha value of 0.05.

Characterization of the Genomic Composition of the SDR

We defined the SDR intervals based on all GWAS loci that passed the Bonferroni correction. SDR intervals were initially defined as +/- 5 kb around each significant GWAS locus. Also, because the SDR region is structurally complex and repetitive, the genome assembly is likely to be inaccurate in this region, thereby reducing the resolution of the GWAS and QTL mapping. We therefore merged all SDR intervals that occurred within 1 Mb on the same chromosome to include all intervening sequence.

To identify levels of polymorphism and divergence between the consensus reference and female-specific haplotypes, we resequenced the F_1 female parent 94006 and the F_2 male parent 'Fish Creek' using 2×250 bp reads on an Illumina HiSeq sequencer. This yielded 106,305,281 paired reads (53 Gb) and 92,077,639 paired reads (46 Gb), respectively. These were aligned to the 94006 reference genome using Bowtie2 with the parameters -D 15 -R 2 -N 0 -L 20 -i S,1,0.75. SNPs were identified using the mpileup function of samtools, followed by bcftools with the parameters -g 1 -O v –m. Since ALLPATHS-LG generates genome assemblies that consist of chimeras of the two haplotypes from a heterozygous diploid genome (Gnerre *et al.* 2011), we expected the *S. purpurea* assembly of chromosome 15 to include segments of Z and W chromosomes. This should be apparent from the relative depth of coverage of female and male sequences. For Z portions of the reference genome, male coverage should be roughly double that of the female for divergent portions of the SDR, whereas for W portions of the reference, coverage should be approximately 0.5X compared to the rest of the genome for the female, and there should be zero coverage in males. We therefore used these alignments to evaluate depth of coverage for the male and female sequences using raw output from the samtools mpileup command to delineate putative Z and W portions of the reference. Similar expectations hold for the GBS markers, which should be homozygous in females and null in males when mapped to the divergent W portions of the reference genome.

We identified female-specific alleles for loci that were heterozygous in the female parent clone 94006 and homozygous in the male offspring, Fish Creek. Sequences containing female-specific polymorphisms (here called "W-type") were created using the FastaAlternateReferenceMaker module of the GATK package (DePristo *et al.* 2011). For comparison, we also used all polymorphisms from the resequencing of both genotypes to create alternative haplotypes using the same approach. Genes with nonsense and frameshift mutations

were then removed as possible pseudogenes. Finally, synonymous (dS) substitution frequencies were estimated for all pairs of predicted transcripts using the '*yn00*' module in the PAML package (Yang 2007). The reference genome transcripts were compared to those containing female-specific polymorphisms as well as to those containing all alternative alleles.

All predicted proteins in the *S. purpurea* reference genome annotation were compared to the UniProt database (http://www.uniprot.org/) using blastp and against the Pfam database (http://pfam.xfam.org/) using HMMER, with default parameters. Protein mapping results were submitted to Argot[2] (Falda *et al.* 2012) to obtain Gene Ontology (GO) annotations, using a stringent cut-off (Total Score=1500) to filter Type I errors. We used Fisher's Exact Test to identify overrepresented GO terms for candidate genes in the SDR. All orthologs between *S. purpurea* and *P. trichocarpa* were retrieved from Phytozome (https://phytozome.jgi.doe.gov/). Synonymous (dS) and nonsynonymous (dN) substitution frequencies were estimated for each pair of primary transcripts from each species using the '*yn00*' module in the PAML package (Yang 2007). Pairs with dS>0.4 were dropped, assuming they were incorrectly defined as orthologs. In total, 33,789 ortholog pairs were compared, including 27,118 genes from *S. purpurea* and 24,000 genes from *P. trichocarpa*.

Estimation of Recombination Rate

As an indicator of recombination rate, we calculated the ratio of physical to genetic distance between marker pairs using linkage groups with >30 markers. For each linkage group, pairwise distances were calculated between every N loci, where N was 10% of the total number of loci on the linkage group. For example, if the linkage group had 100 markers, the distance was calculated between all pairs of loci that were separated by 10 loci. Negative and extreme values (ratio>15) were removed for the purpose of visualization.

Gene Expression

RNA sequencing was performed for actively growing shoot tips for five male and five female progeny from the family used for QTL analysis. Detailed methods are described in Carlson et al. (2017). Briefly, total RNA was extracted using the Spectrum™ Total Plant RNA Kit. Libraries were constructed using the NEBNext Ultra Directional RNA Library Prep Kit. Libraries were sequenced on the Illumina HiSeq platform (1x100 bp) yielding an average of 17.9 million mapped reads per sample. Reads were mapped to the *S. purpurea* reference genome v1.0

using the CLC Genomics Workbench, and differential expression analyses were performed using EdgeR.

Results

Localization of the SDR to Chromosome 15 and Chromosome 19

Among the 396 phenotyped and genotyped individuals in the F_2 family, there were 234 females and 162 males. This ratio is significantly skewed toward females (F:M=1.44; χ^2=13.1; df=1; P<0.001). QTL mapping identified sex-associated markers principally on chromosome 15 for all three maps (Figure 1; Table S1). On the female map, 125 markers were linked to sex, 105 of which were on chromosome 15, spanning from 225.42 cM to 240.17 cM (Table 1). On the male map, only five markers were linked to sex, four of which were in the interval from 326.48 cM to 347.17 cM on chromosome 15 (Figure 1, Table 1). An additional 50 markers were linked to sex on the intercross map, covering an interval of about 2.6 cM, all on chromosome 15 (Figure 1, Table 1). Based on anchoring mapped markers to physical positions in the *S. purpurea* genome assembly, the potential SDR can be mapped to two regions on chromosome 15 ranging from ~0.4 Mbp to 1.9 Mbp and from ~10.9 Mbp to ~15.1 Mbp.

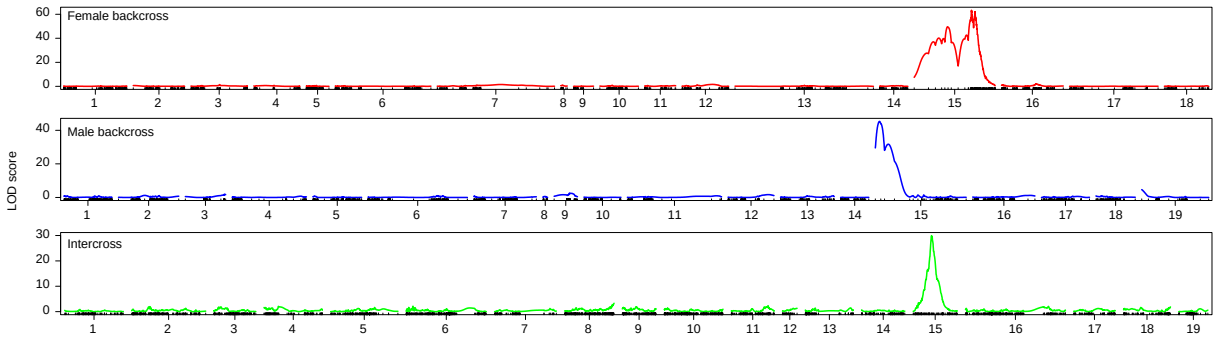

Figure 1 QTL for sex in an F_2 *S. purpurea* cross. From top to bottom are LOD scans for female backcross (red), male backcross (blue) and intercross (green) markers across the 19 major *S. purprea* linkage groups. Chromosome 15 has a very strong QTL sex in all three maps, and the female backcross also shows a weak peak on chromosome 19 (LOD=4.68; table 1).

Table 1 Bayesian credible intervals for sex QTL on chromosome 15.

	Physical Map		Genetic Map	
	Start (bp)	End (bp)	Start (cM)	End (cM)
Female Map	10,939,613	11,569,298	225.42	240.17
Male Map	372,445	1,881,243	326.48	347.17
Intercross	11,401,384	15,091,498	55.69	58.22

One additional sex-linked marker was located at the proximal end of chromosome 19 on the male map, with a LOD score of 4.68 (Figure 1; Table S1). However, mapping failed entirely for chromosome 19 for female backcross markers, the only chromosome for which this was the case. Chromosome 19 had the lowest density of GBS markers in the genome (Table S2). Furthermore, this chromosome had the lowest proportion of markers in a female-backcross configuration, and the highest proportion of markers with severe segregation distortion (Figure S2; Table S2).

To confirm the location of the SDR in a diverse population, a GWAS for sex was performed using naturalized *S. purpurea* accessions collected from northeastern North America. Of the 60 genets that were unambiguously phenotyped for sex, 38 were female and 22 were male, which is a significantly female-biased sex ratio (F:M=1.73; χ^2=4.3; df=1; *P*=0.02). Of the 85,543 SNP markers that passed filtering, 72 were significantly associated with sex ($P<5.85 \times 10^{-7}$, Figure 2; Figure S3). Among these markers, 41 were located on chromosome 15, from 10.7 Mb to 15.3 Mb, and four were located at the distal portion of chromosome 15 (1.9 Mbp). Thus, the primary SDR identified by GWAS overlaps with those mapped by QTL in the F_2 family (Figure 3). In addition, six markers from chromosome 19 at ~69 kb were also significantly associated with sex (Figure 2), which also corresponds with the QTL results. Additionally, there were minor peaks on chromosomes 1,2,3, and 5, and there were six scaffolds containing a total of 13 significant sex-associated markers that were not anchored to the genetic maps (Table S3).

To evaluate whether these secondary chromosomal peaks could have been due to assembly errors, we aligned these SDR sequences to the *S. purpurea* reference genome using blastn. None of these chromosomal loci shared homology with the chromosome 15 SDR (Table S4). We also compared these SDR sequences to the *Populus trichocarpa* v3.0 reference genome using blastn. The SDRs on chromosomes 1,2, and 5 had best hits to the same chromosomes in *P. trichocarpa*. However, the SDRs on chromosomes 3 and 19 had best hits to scaffold_25 in *P.*

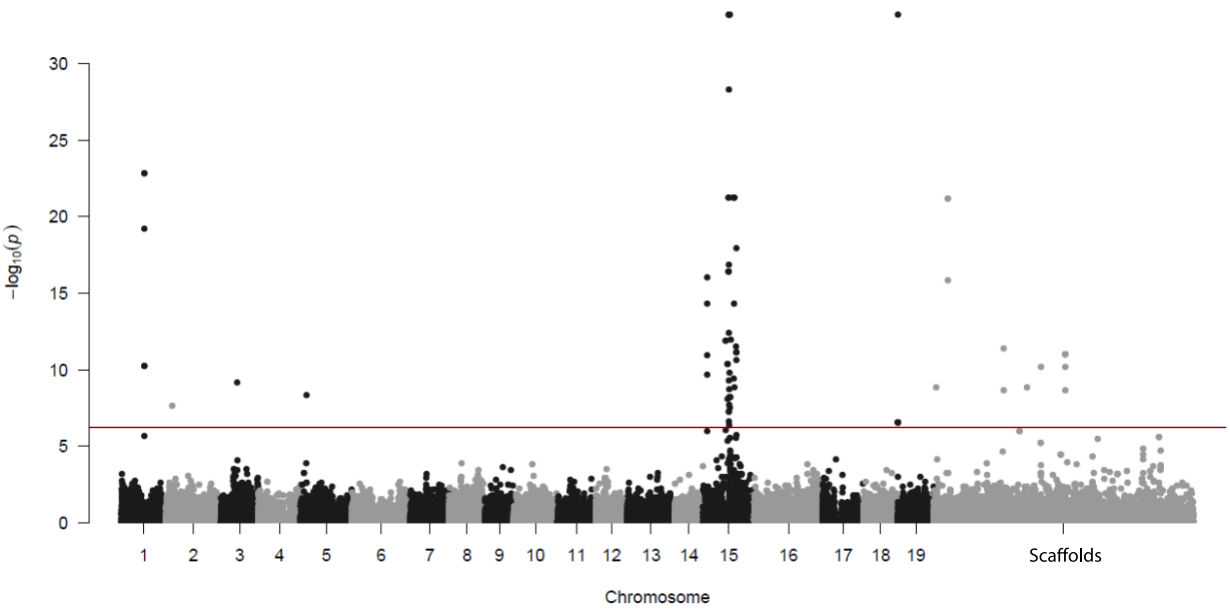

Figure 2 Manhattan plot derived from genome-wide association analysis for sex determination. The Y-axis shows the strength of association ($-\log_{10}$(P value)) for each SNP ordered by chromosome and SNP position (x axis). The horizontal line indicates significance after a Bonferroni correction for multiple testing.

trichocarpa (Table S4). Because the SDR is known to be poorly assembled in the *P. trichocarpa* v3.0 assembly (Geraldes *et al.* 2015), we aligned scaffold_25 to the *P. trichocarpa* v1.0 assembly and found that it matched primarily to chromosome 19, positions 751 to 1040 kb, which coincides with the main *P. trichocarpa* SDR (Geraldes *et al.* 2015). Therefore, the QTL and GWAS results both indicate that sequences homologous to the *P. trichocarpa* SDR retain evidence of sex dimorphism in *S. purpurea*.

S. purpurea Has a ZW System of Sex Determination

Under Mendelian segregation, the frequency of heterozygotes should be 0.5 for both male and female F_2 progeny. However, the frequency of heterozygosity was 0.64 for female progeny and only 0.12 for males (Table S1). The skewed heterozygosity occurred in blocks in the vicinity of the sex QTL peaks (Figure 3). Furthermore, females in the association population had an average observed heterozygosity of 0.79 for the sex-associated SNP loci, while males had an observed heterozygosity of only 0.05 for these same loci (Figure 4a, Table S3, Figure S4). This difference was significant based on a t-test ($P < 2.2\text{x}10^{-16}$). Both observations are consistent with a female heterogametic (ZW) system of sex determination, where females should be nearly fixed heterozygous for female-specific portions of the SDR, while males should be homozygous for

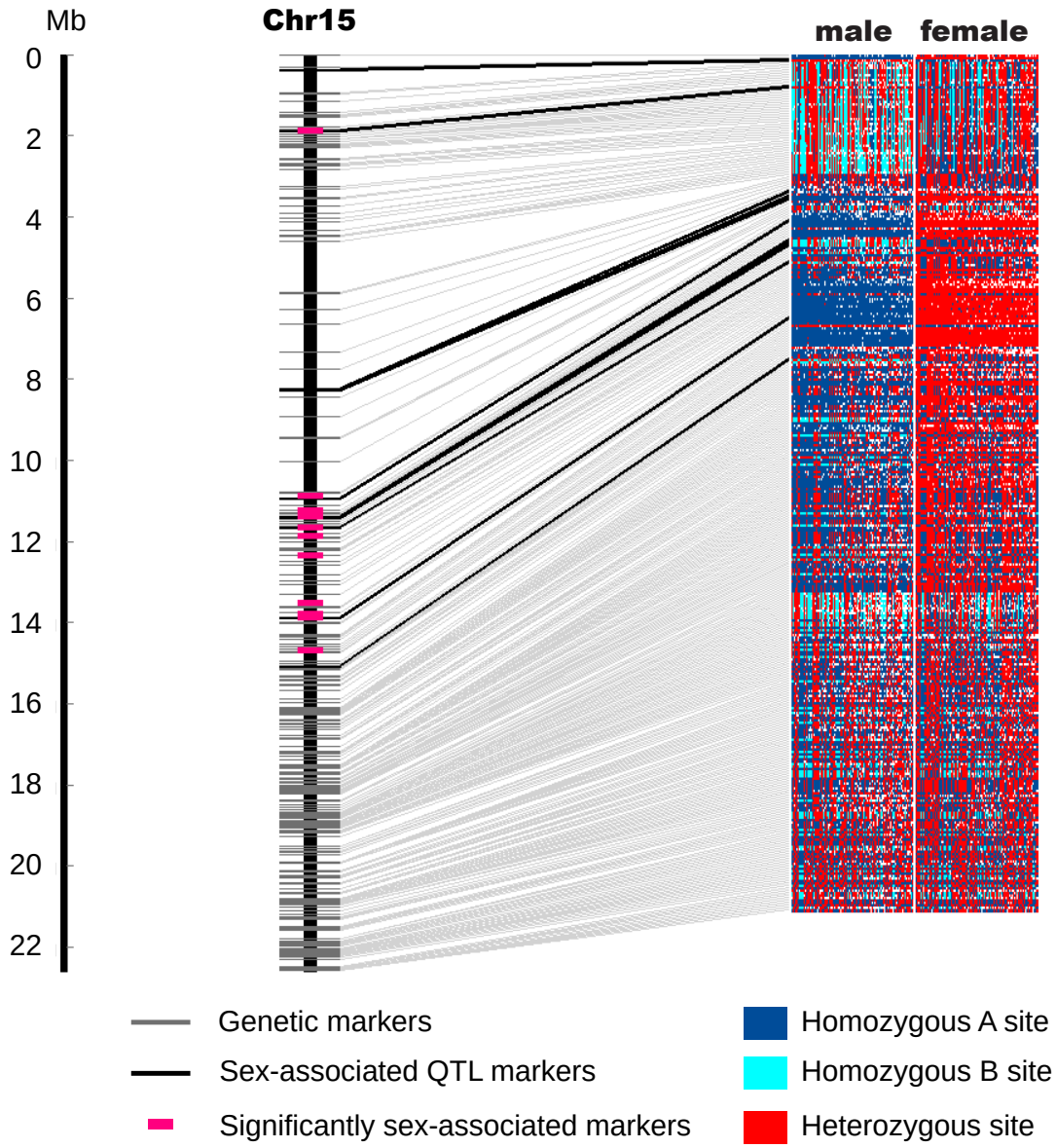

Figure 3 Genotype configurations in males and females from the F_2 family on chromosome 15. Markers from all three genetic maps are shown as horizontal lines corresponding to their physical positions on the chromosome 15 physical assembly. Markers with top LOD scores in each map are colored as black. Significantly associated markers from the GWAS analysis with $P < 1x10^{-7}$ are indicated by fuschia marks on the physical map. Each marker is connected between physical map and its genotype configurations with 100 selected progeny of each sex. Genotypes of QTL markers are colored according to their homozygosity or heterozygosity.

323

those same loci. This is due to the typically biallelic nature of SNP polymorphisms, where polymorphic alleles from the W chromosome are identical by descent and therefore only occur in females. The discrepancy between the observed values and the expected fixed heterozygosity in females is likely due to null alleles caused by allele dropout and/or inadequate sequencing depth for the GBS markers (Andrews *et al.* 2016).

Since our reference sequence was derived from a female, we expected that the assembly could contain hemizygous or highly divergent portions of the W chromosome. We used two complementary approaches to determine the size and extent of these regions: allelic configurations of the GBS markers, and relative depth of sequence coverage in male and female clones. Candidate W segments contained a large proportion of GBS markers that were homozygous in females and mostly lacking genotype calls (i.e., double null markers) in males in the association population (Figure S5). We identified 231 of these W-type markers (0.27%) (Figure 4a; Table S5). Of these, 51 occurred on chromosome 15, another 158 occurred on 20 unanchored scaffolds, and the remaining 22 occurred on small segments of chromosomes 3, 5, and 7. The genotype configuration for these markers was consistent with a ZW system, such that the average observed homozygosity for females was 0.80 (presumably due to hemizygosity or divergence of W segments) whereas 85% of males had null alleles at these loci, on average (Figure 4a, Table S5). The putative W haplotypes were interspersed along chromosome 15, suggesting that the genome assembly is a chimeric representation of the Z and W haplotypes (Figure 4a; Table S5).

We also examined depth of coverage in male and female reference-based assemblies to identify putative hemizygous W chromosome segments in the reference genome. If females are heterogametic, then there should be regions in the female reference that are not covered by reads from a male individual. Aligning paired 250 bp Illumina sequences from a male offspring ('Fish Creek') of clone 94006 back to the female reference assembly, yielded a very high alignment rate of 95.19% compared to 96.67% when clone 94006 was aligned to itself. Nevertheless, after excluding known repeats and gaps, there were 22,733 regions totaling 7.69 Mb on chromosomes and another 6.87 Mb of unanchored scaffolds that had coverage in the female but lacked coverage in the male (Table 2; Figure S6). These analyses identified 222 scaffolds comprised of >30% female-specific sequences (Table S5). Some of these are likely caused by insertion/deletion polymorphisms that are not sex-specific. However, we identified 11 scaffolds

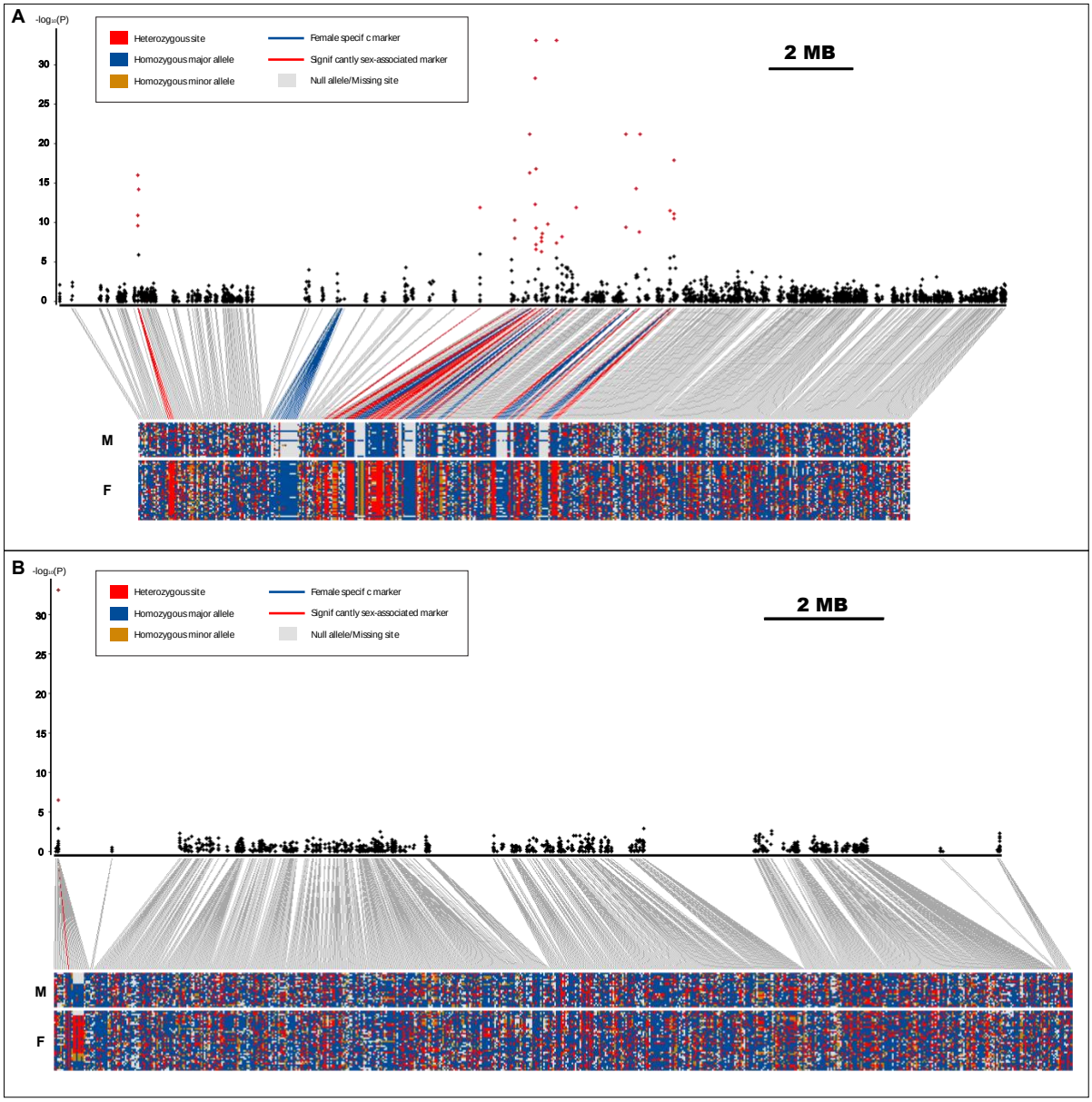

Figure 4 Genotype configurations of markers on chromosome 15 (A) and chromosome 19 (B) from the *S. purpurea* association population. The top is a blowup of chromosome 15 from the Manhattan plot in Figure 2, with significantly sex-associated markers colored red. The bottom shows the genotype configurations of 22 males and 38 females from the association population, where each row represents an individual. "Major alleles" are here defined as those with higher frequency in males, and are shaded blue where homozygous, while homozygotes for male minor alleles are shaded gold. Heterozygous sites are shaded red, and missing data is light gray. Lines connect each plotted marker to its physical position. Red lines indicate that markers are significantly associated with sex while blue lines indicate the markers were identified as female-specific (putatively derived from the W haplotype).

that were also identified as putative W segments based on allelic configurations (see above). Portions of five of these scaffolds had high sequence similarity to chromosome 15, supporting the contention that these are alternate haplotypes from the SDR. For example, Scaffold0265 is 298 kb in length and contains 38.9% female-specific sequence and 20 W-type GBS markers (Table S6). This scaffold also contains three sex-associated markers identified in the GWAS. Cumulatively, these 11 scaffolds covered 1.04 Mb, which is a reasonable lower limit for the size of the divergent portions of the SDR.

Table 2 Length of intervals that lacked coverage in alignments of 2x250 bp reads against the reference genome assembly (also derived from female clone 94006). Number in the parentheses is the percentage of the total genome composition in that category that lacked coverage.

	Whole Genome	**Fish Creek (♂)**	**94006 (♀)**
Total Length	348,745,509	14,564,089 (4.18)	562,813 (0.16)
Chromosomes	251,661,964	7,693,428 (3.06)	303,356 (0.12)
Scaffolds	97,083,545	6,870,661 (7.08)	259,457 (0.27)
Repeats	98,506,863	5,328,429 (5.41)	260,598 (0.26)
Genes	120,852,638	2,654,305 (2.20)	78,325 (0.06)
SDR	3,073,122	480,360 (15.63)	4,814 (0.16)

The SDR is Highly Repetitive, Has Repressed Recombination, and is Divergent from the *Populus* SDR

The SDR on chromosome 15 of *S. purpurea* overlaps with a large region (9.8 Mb to 16.2 Mb) with elevated physical-to-genetic distance ratio of 0.867 Mb/cM, compared to the genome-wide average of 0.172 Mb/cM (Figure 5), which indicates reduced recombination. This interval contained high repeat abundance relative to the rest of the genome (Figure S7). A portion of the SDR in *S. purpurea* is homologous to the SDR in *S. suchowensis*. The *S. suchowensis* SDR primarily occurs on scaffold64, an ~900 kb scaffold that maps to chromosome 15 (Hou *et al.* 2015). Aligning this sequence to the *S. purpurea* genome with lastz, we observed homology from 6.2 to 7.3 Mb and from 14.1 and 15.1 Mb on *S. purpurea* chromosome 15 (Figure S8). The latter sequence overlaps with a portion of the *S. purpurea* SDR. In contrast, the *S. viminalis* SDR

matches from 5.9 to 8.4 Mb on *S. purpurea* chromosome 15, which is outside the *S. purpurea* SDR (Pucholt, Wright, *et al.* 2017).

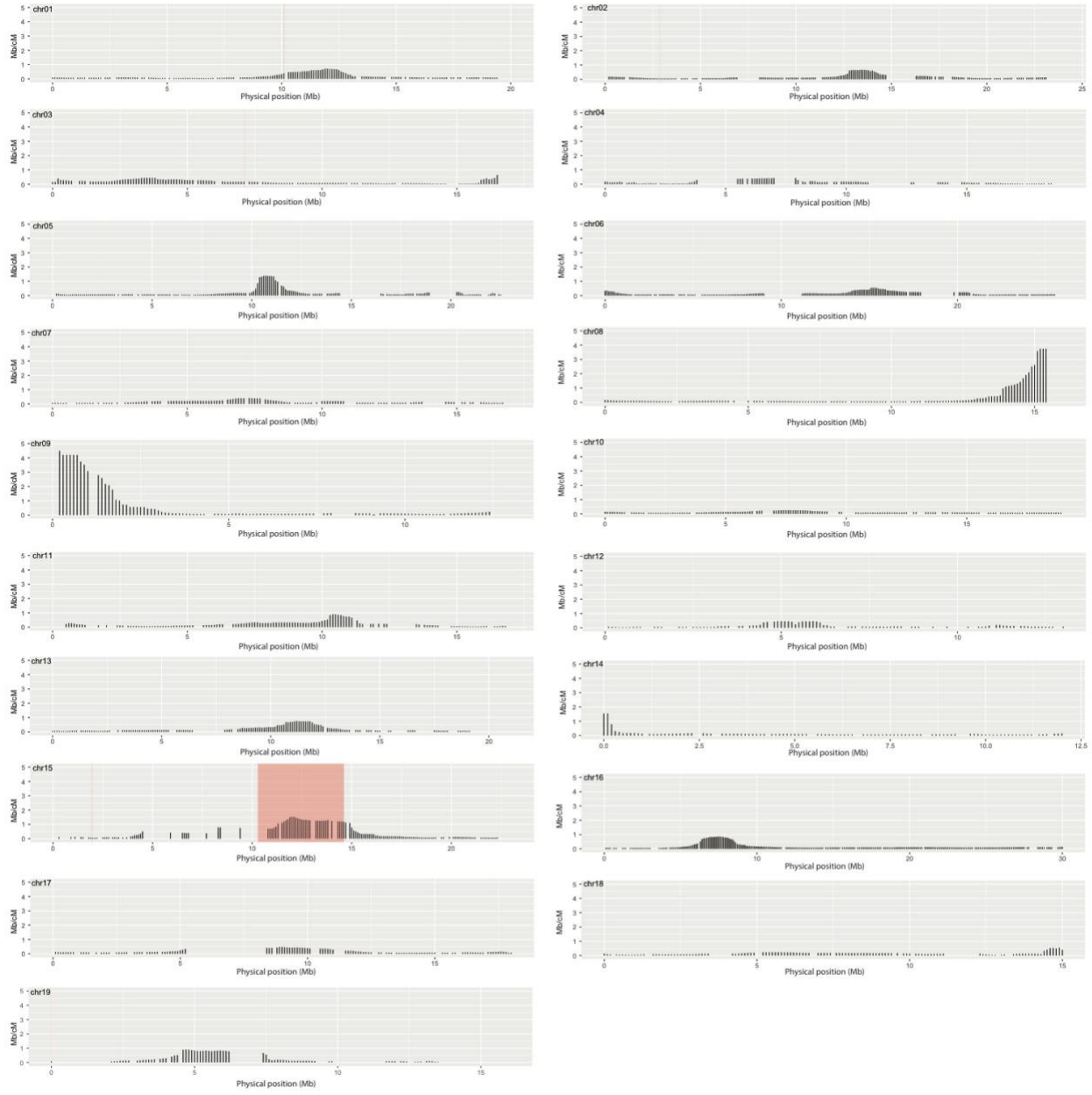

Figure 5 Recombination across the *S. purpurea* genome, as inferred from physical:genetic distance ratio. Bar plots represent the physical:genetic distance ratio (Mb/cM) in 100 kb windows for the 19 chromosomes. The position of the SDRs are indicated by vertical red shading.

P. trichocarpa is another member of the Salicaceae and has a fairly-well characterized XY system of sex determination (Geraldes *et al.* 2015). In general, *S. purpurea* and *P. trichocarpa* have high synteny at the chromosome scale (Figure 6), but chromosome 15 in *S. purpurea* stands out in several ways. First, the SDR on chromosome 15 of *S. purpurea* is not syntenic with chromosome 15 or any other chromosome of *P. trichocarpa* (Figure 6). Second, the proportion of repeats is significantly elevated in the *S. purpurea* SDR, with an average of 37% repeat composition, compared to the genome-wide average of 24.8% (Welch's Two-Sample T = -4.6 *P*5948, <0.0001; Table S7; Figure S7). Chromosome 19, which contains the SDR in *P. trichocarpa*, also had the highest average repeat content in *S. purpurea* (33.5%, compared to 25.1% genome-wide average) (Table S7).

Gene Content of the SDR

We identified 251 protein-coding genes within the *S. purpurea* SDR (Table S8). A GO enrichment analysis based on 203 genes annotated with GO terms identified 4 significantly enriched terms (Bonferroni adjusted $P < 2.45 \times 10^{-4}$), all of which were related to microtubule functions. These include microtubule-based movement (GO:0007018), microtubule motor activity (GO:0003777) and microtubule binding (GO:0008017), as well as kinesin complex (GO:0005871) (Table 3). This enrichment is partly due to two pairs of tandemly-duplicated kinesin-like genes in the SDR (Table S8). Since there is only one homolog of these kinesin-like genes in *P. trichocarpa*, it appears that this expansion occurred after the divergence of the two genera, a scenario supported by high sequence conservation between the tandem duplicates (Figure S9).

Table 3 Significantly overrepresented GO terms of candidate genes from SDR.

Description	GO term	Number of genes in SDR	Number of genes outside SDR	P value
Microtubule motor activity	GO:0003777	7	91	4.73×10^{-6}
Kinesin complex	GO:0005871	7	92	5.07×10^{-6}
Microtubule-based movement	GO:0007018	7	92	5.07×10^{-6}
Microtubule binding	GO:0008017	7	133	4.84×10^{-5}

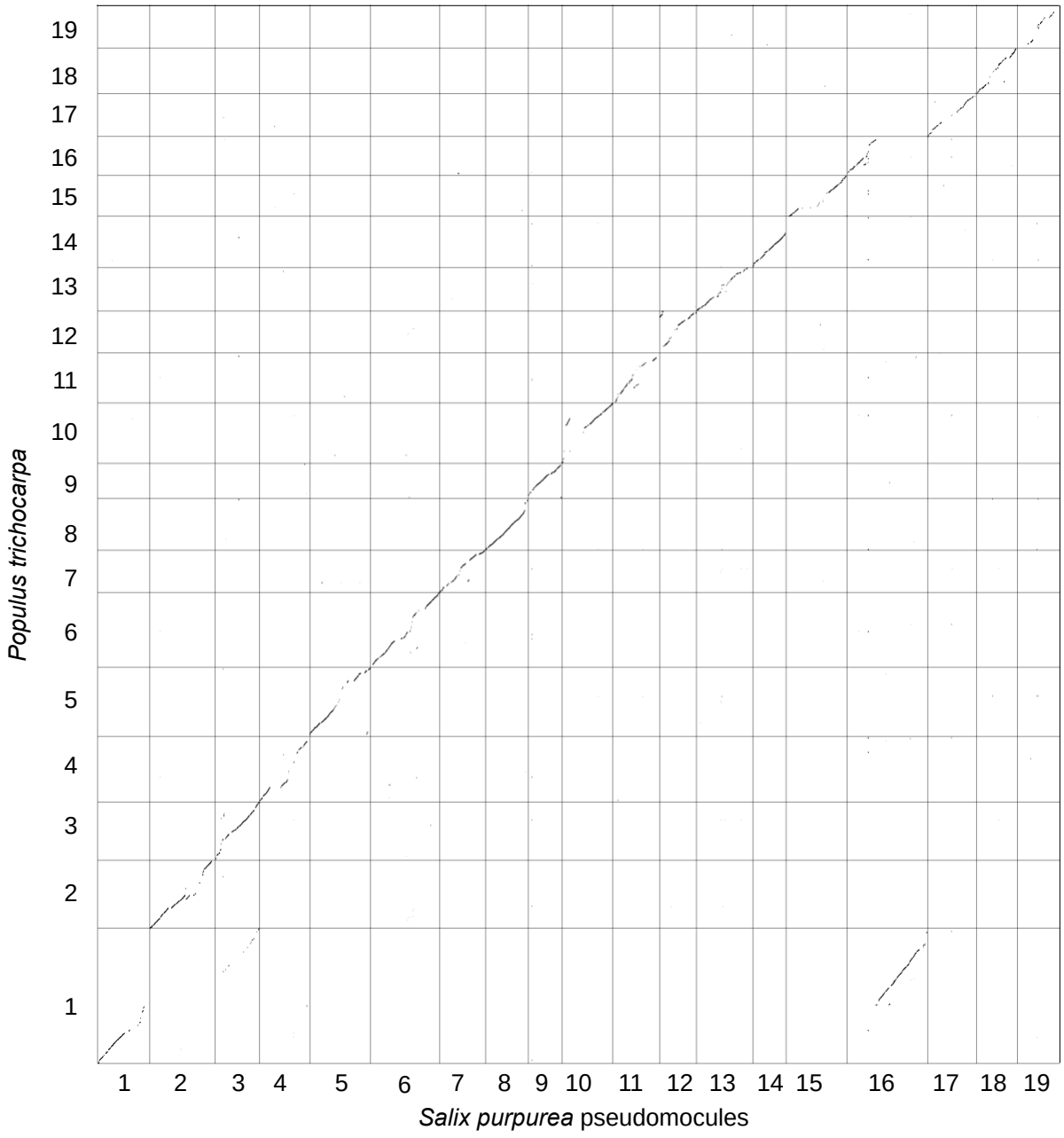

Figure 6 Comparison between the *S. purpurea* (x-axis) and *P. trichocarpa* (y-axis) genomes. The two genomes are largely syntenic based on genome-scale alignments using LASTZ, with parameters set to exclude paralogous segments derived from the most recent whole genome duplication.

409

The SDR contains 20 genes that have >70% female-specific sequence, and many of these genes also show sex-biased expression in stem tissue in *S. purpurea* (Table S8; Carlson *et al.* 2017). These include an extracellular calcium-sensing receptor (SapurV1A.0301s0080), an auxin response factor (SapurV1A.0718s0100), a peptidase M50B-like protein (SapurV1A.0475s0170), a zinc finger C3hC4 type transcription factor (SapurV1A.0301s0170), and a reticulon-like protein (SapurV1A.0530s0130). Among these, only the reticulon-like protein showed an elevated dN/dS ratio when compared to *P. trichocarpa* (0.687, versus a genome-wide average of 0.406). Of the 14 genes that showed significant female-biased expression in the SDR, only one lacked female-specific sequence (SapurV1A.1386s0030, a small heat shock protein). No genes showed significant male-biased expression after Bonferroni correction.

The chromosome 19 SDR is of particular interest, since it overlaps with the SDR of *P. trichocarpa*. This region spans approximately 10 kb in the current assembly, and harbors three small genes. SapurV1A.1005s0060 contains a Small MutS-Related (SMR) domain. A second gene, SapurV1A.1005s0050, is a calcium-dependent kinase with two EF-Hand domains. The third gene, SapurV1A.1005s0070, encodes a hypothetical protein (Table S8). None of these genes have sex-biased expression or unusual dN/dS ratios compared to *Populus* (Table S8).

We attempted to estimate the relative age of the region of suppressed recombination by estimating the rate of synonymous substitution of W alleles compared to Z alleles in the SDR. Calculated this way, the Z-W synonymous substitution rate within the SDR was 0.00343 while the rate calculated the same way outside of the SDR was 0.00151. These differences were statistically significant (t = -4.099; df = 249; P = 5.63e-05). For comparison, we also calculated divergence between alleles using all observed polymorphisms. Genes within the SDRs showed similar overall divergence (dS=0.00616) compared to genes outside the SDRs (dS=0.00607), and the difference was not significant (t = -0.077; df = 235; P = 0.938). There was no evidence of evolutionary strata in the SDR based on lack of clustering of genes with similar dS values.

Discussion

The *S. purpurea* SDR is Similar to Other *Salix* Species and Divergent from *Populus*

In all three of the *Salix* species studied thus far, *S. viminalis* (Pucholt *et al.* 2015), *S. suchowensis* (Hou *et al.* 2015; Chen *et al.* 2016), and now *S. purpurea*, the largest SDR is on

chromosome 15, and shows clear female heterogamety. Furthermore, the *S. suchowensis* SDR overlaps with a portion of the *S. purpurea* SDR, but the *S. viminalis* SDR does not. This may reflect the evolutionary distinctness of *S. viminalis* from the other two taxa. Based on morphological characters, *S. viminalis* belongs to section *Viminella*, which is strongly differentiated from section *Helix*, which contains *S. purpurea* (Argus 1997). This is similar to the situation in *Populus*, where the location of the sex determination region varies across different sections of the genus, though all are located on chromosome 19 (Gaudet *et al.* 2008; Pakull *et al.* 2009, 2014; Paolucci *et al.* 2010; Tuskan *et al.* 2012; Kersten *et al.* 2014; Geraldes *et al.* 2015). Comparison of the sequence composition of the *Salix* SDRs and the *P. trichocarpa* SDR revealed no extensive stretches of homology, suggesting a largely independent evolution of these genome regions (Hou *et al.* 2015; Pucholt, Hallingbäck, *et al.* 2017). Clearly, the SDR is highly dynamic within this family.

The alternative peaks from the GWAS analysis on chromosomes 1, 2, 3, and 5 were not upheld by the QTL analysis, and mainly consisted of isolated markers. This is unlikely to represent a case of multi-locus sex determination (Moore and Roberts 2013), as the evidence is weak since there is little other corroborating information. The peaks on chromosomes 2, 3, and 5 consisted of solitary markers, while that on chromosome 1 included 5 markers that occurred within a 1 kb interval. Our results are similar to those in *P. trichocarpa*, which also contained multiple secondary GWAS peaks in a sex determination GWAS (Geraldes *et al.* 2015). While some of the secondary *Populus* peaks appear to be assembly and/or alignment artifacts (Geraldes *et al.* 2015), we found no evidence of assembly errors in these regions for *S. purpurea* based on examining the sequence assembly itself as well as the underlying genetic map. Problems with assembly of SDRs are common, presumably due to strong haplotype divergence and high repeat composition, which impede assembly of short-read sequence data (Miller *et al.* 2010). Furthermore, the suppressed recombination in these regions inhibits map-based assembly methods. An alternative explanation for the secondary peaks is recent translocation from those chromosomes to the W chromosome in *S. purpurea*. If the W haplotype is not represented in the reference genome assembly, then the reads derived from the recently-translocated regions could align to their original locations. Short-read sequence aligners like Bowtie2 do not handle repetitive sequences well, and commonly misalign reads derived from such regions (Lian *et al.* 2016).

The GWAS peak on chromosome 19 are especially interesting because it coincides with the position of one of the SDRs in *Populus*. This peak also has more corroborating evidence than the other secondary peaks because it had one of the lowest observed P-values, and it is recapitulated in the QTL analysis. Furthermore, the peak on chromosome 3 best matches a scaffold from the SDR region of *Populus* on chromosome 19, so at least two independent association results point to sex-specific genotypes in genomic segments with homology to the *Populus* SDR. If these represent recent translocations, then this could be a clue to the origin of the chromosome 15 SDR in the *Salix* lineage.

Recombination Suppression and Relative Age of the SDR

Reduced recombination is a crucial component of sex chromosome evolution which ensures that male and female sterility factors do not co-occur in the zygote (Bergero and Charlesworth 2009; Ming *et al.* 2011). As expected, we observed reduced recombination across most of the SDR in *S. purpurea* (Figure 5). This could be caused by large-scale structural polymorphisms and reinforced by the accumulation of nonhomologous sequences in the female-specific haplotype (Ming *et al.* 2011; Charlesworth 2015). The SDR also shows a higher proportion of repetitive elements, as expected in regions with reduced recombination. Similar features are also apparent within the SDR of *S. suchowensis* and *S. viminalis* (Hou *et al.* 2015; Pucholt *et al.* 2015; Chen *et al.* 2016), but are not as apparent for the *P. trichocarpa* SDR, which is estimated to be quite small (Geraldes *et al.* 2015). If this is accurate, it could indicate that the *P. trichocarpa* region has not yet developed these features, or that it is highly dynamic. In the case of *S. purpurea*, the SDR is quite large, with a lower limit of 1.04 Mb (based on the cumulative length of female-specific scaffolds), and an upper limit of approximately 5 Mb, based on suppressed recombination and the occurrence of SNPs that are significantly associated with sex. It is possible that the SDR overlaps with the centromere on chromosome 15, and this could contribute to the large apparent size of the region of suppressed recombination. However, the SDR does not contain any of the tandem minisatellite repeats that are apparently characteristic of the *S. purpurea* centromeres, as identified in a previous study (Melters *et al.* 2013). It remains to be seen if the lack of these repeats is due to poor assembly, or if the centromere is located elsewhere on this chromosome.

Divergence between Z and W transcripts in the *S. purpurea* SDR is relatively low, suggesting that suppression of recombination is incomplete or recently established. This is

similar to the SDRs of *P. trichocarpa* (Geraldes *et al.* 2015) and *S. viminalis* (Pucholt, Wright, *et al.* 2017), which also show low divergence of sex-specific sequences. Furthermore, we saw no evidence of the presence of evolutionary strata within or around the *S. purpurea* SDR. Such features occur due to the establishment of regions of suppressed recombination at different times during sex chromosome evolution (Charlesworth 2016). Evolutionary strata are apparent in well-established SDRs of other plants, including *Silene latifolia* (Bergero *et al.* 2007) and *Carica papaya* (Wang *et al.* 2012). However, no such regions were detected in *S. suchowensis* (Pandey and Azad 2016). Given the low divergence, lack of strata, and the frequent movement of the SDR within the family, it is reasonable to conclude that the SDR is highly dynamic in this family, and that sex determination loci frequently translocate to new positions and/or are superseded by other loci on autosomes, as predicted by theoretical models of SDR movement (van Doorn and Kirkpatrick 2007, 2010).

Candidate Genes and Their Function

The SDRs are genomic regions that are statistically associated with gender. This association must be due to the presence of loci that control sex determination, but the regions also likely harbor loci that are under sexually antagonistic selection (van Doorn and Kirkpatrick 2007; Bachtrog *et al.* 2014). The gene content of these regions could therefore provide insights about mechanisms of sex determination as well as sex dimorphism. We identified 251 protein-coding genes in the SDRs of *S. purpurea* (Table S8). Most have not been functionally annotated, but clues can be inferred based on conserved domains and their predicted function in model organisms. It is also important to note that the assembly problems mentioned previously have probably prevented full enumeration of the gene content of the SDRs. This problem may be particularly challenging for female-specific portions of the W chromosome (Pucholt *et al.* 2015). Nevertheless, there are several genes in this region that could plausibly be involved in floral development and sex-specific regulation that are worthy of consideration.

Since floral morphology is the most striking difference between the sexes, it is reasonable to expect that genes involved in floral development would be located in the SDRs. Indeed, the SDR contains SapurV1A.0718s0010, an ortholog of WUSCHEL-related homeotic genes (e.g., *WOX1*). Orthologs in other species, including STF in *Medicago truncatula*, LAM1 in *Nicotiana sylvestris*, and MAW in *Petunia*, are key regulators of the lateral outgrowth of leaf blades and

floral organs (Lin *et al.* 2013). This gene showed slightly elevated expression in male shoot tips compared to female shoot tips (Table S7).

Several genes in the SDR may be involved specifically with male development and function. For example, our analysis of GO term over-representation highlighted the presence of seven genes containing the kinesin motor domain (PF00225), which is involved in microtubule-based movement or organelles, including during pollen tube growth (Cai and Cresti 2009). For example, loss-of-function mutants of the closest homolog of SapurV1A.0530s0110 in *Arabidopsis thaliana* (*NACK*1) showed reduced growth and prematurely-terminated petals, pistils, and stamens (Nishihama *et al.* 2002).

Two other genes in the SDR may be related to pollen function. First, SapurV1A.1741s0030, is a homolog of *PLANT INTRACELLULAR RAS GROUP-RELATED LRR 3* (*PIRL3*), which has been implicated in pollen development in *Arabidopsis* (Forsthoefel *et al.* 2013). The second, SapurV1A.0301s0130 is a homolog of the *Arabidopsis* gene AT3G01570.1, a member of the oleosin family (Kim *et al.* 2002). This gene occurs in a female-specific portion of the SDR and has female-biased gene expression, perhaps reflecting a potential role in sex dimorphism in *Salix*. Other members of the oleosin family have been associated with pollen wall development, and eight of these have been shown to have tapetal-specific expression in *Arabidopsis* (Kim *et al.* 2002; Hsieh and Huang 2004; Yang *et al.* 2007). The expression of one of these oleosin genes is regulated by *MALE STERILITY1* (*MS1*) in *Arabidopsis*, which controls pollen and tapetal development (Yang *et al.* 2007), raising the possibility that altered regulation of oleosin genes could provide a pathway to male sterility.

The SDR on chromosome 19 deserves special attention due to its shared homology with the *Populus* SDR. One particularly interesting gene in this region is SapurV1A.1005s0060, which contains a Small MutS-Related (SMR) domain and a domain of unknown function (DUF1771). These domains frequently occur together in eukaryotes, but the function of DUF1771 has yet to be characterized (Fukui and Kuramitsu 2011). Proteins with the SMR domain, such as MutS2, can suppress (Fukui *et al.* 2007; Fukui and Kuramitsu 2011) or promote (Burby and Simmons 2017) homologous recombination by endonucleolytic digestion, and are involved in mismatch repair in diverse prokaryotes (Kunkel and Erie 2005). The roles of the SMR domain in plants are not fully characterized, but when coupled with the pentatricopeptide repeat motif, the SMR domain shows sequence-specific RNA endonuclease activity and affects

chloroplast function (Zhou *et al.* 2017). Due to its potential roles in recombination, mismatch repair, and regulation of organellar function, this gene is an intriguing candidate in the context of sex determination as well as mediation of the female-biased sex ratios that are commonly observed in *Salix* (Alliende and Harper 1989; Alstrom-Rapaport *et al.* 1998; Ueno *et al.* 2007; Pucholt, Hallingbäck, *et al.* 2017), including in *S. purpurea*, as reported here.

Sex Chromosome Evolution in the Salicaceae

Populus and *Salix* are closely-related genera that share many key characteristics, the most notable of which is that they are both nearly fixed for dioecy. *Populus* first appears in the fossil record between 40 and 60 MYA, apparently slightly earlier than *Salix* (Boucher *et al.* 2003). However, *Populus* and *Salix* exhibit much less divergence in nucleotide sequence and chromosome structure than expected, presumably due to long average generation times (Sterck *et al.* 2005; Hou *et al.* 2016). It may therefore seem surprising that the chromosomal location and gene content of the SDRs are so different, and that they have different heterogametic configurations (Hou *et al.* 2015; Pucholt *et al.* 2015). In fact, movement of sex determination loci and transitions between XY and ZW systems are well-known in organisms that lack strongly-differentiated, heteromorphic sex chromosomes (Bachtrog *et al.* 2014).

A striking finding of this study is the existence of multiple sexually dimorphic regions in the *S. purpurea* genome, one of which is on chromosome 15 and shared with other *Salix* species (Pucholt *et al.* 2015; Chen *et al.* 2016), and one on chromosome 19, which harbors the SDR of multiple *Populus* species (Tuskan *et al.* 2012; Kersten *et al.* 2014; Geraldes *et al.* 2015). There are several lines of evidence to support a model whereby the original sex determination locus was located on chromosome 19 in the common ancestor of *Salix* and *Populus*. First, *Salix* chromosome 19 shows overall high synteny with *P. trichocarpa* chromosome 19, and both species apparently have an SDR in the same region of this chromosome. In contrast, chromosome 15 shows no sex chromosome characteristics in *Populus*, and the composition of chromosome 15 is quite different in the *Salix* SDR. Second, we were unable to create a female backcross map for *S. purpurea* chromosome 19, the only chromosome for which mapping failed. This was due to a paucity of genetic markers in this region, particularly in the female backcross configuration. Notably, such a configuration would be absent in the SDR of an XY sex determination system. On the other hand, the overall lack of genetic markers is probably caused in part by the high repeat content of this chromosome, which can inhibit genotyping based on

short read sequences (Treangen and Salzberg 2013). High repeat content is also expected in regions of reduced recombination, commonly found in sex chromosomes (Ming and Moore 2007; Bergero and Charlesworth 2009). In summary, because chromosome 19 shows characteristics of an SDR in *S. purpurea* and in multiple *Populus* species, but chromosome 15 only shows such characteristics in *Salix*, it is logical to hypothesize that chromosome 19 is the ancestral sex chromosome in the Salicoid lineage.

In the present study, it is important to reemphasize that the locus mapped to chromosome 19 may be an assembly or alignment artifact. This could be caused by a recent translocation from chromosome 19 to the W haplotype of chromosome 15, which would result in incorrect alignment of GBS reads to the original chromosome 19 locus if the W haplotype is not in the main genome assembly. However, because the locus matches a portion of the SDR of chromosome 19 in *Populus*, and the gene content of these regions is similar between the taxa, this finding would still provide valuable clues about sex determination and/or sex dimorphism in this family even if it is caused by a recent translocation. It is also noteworthy that the *S. purpurea de novo* genome assembly did not use the *P. trichocarpa* genome assembly as a reference to guide placement of scaffolds in pseudomolecules (Smart et al., in preparation), so the results reported here are not caused by carryover of biases or errors from the original *P. trichocarpa* assembly.

Unfortunately, a definitive comparison of the Salicaceae sex chromosomes is not possible with the currently-available genome sequences. The SDRs of *Salix* and *Populus* are typical in that they have complex structural polymorphisms, high repeat content, and low recombination rates, all of which contribute to fragmentary and erroneous genome assemblies (Geraldes *et al.* 2015). Efforts are underway to assemble these regions using long read sequencing and dense genetic mapping in multiple pedigrees. This will facilitate analyses that can date the origin of these regions based on differentiation of sex-specific haplotypes in the non-recombining portions of the SDR (Otto *et al.* 2011). Furthermore, elucidation of the sex determination system in additional Salicaceae taxa should help to determine the ancestral state. This family should therefore be instrumental in advancing our knowledge of the evolution and ecological significance of sex chromosomes as genetic and genomic resources continue to accumulate.

Acknowledgements

We are grateful to Matt Olson and two anonymous reviewers for helpful comments on the manuscript. This work was supported by grants from the USDA-NIFA CAP program (4705-WVU-USDA-9703), the DOE JGI Community Sequencing Program, and the NSF Dimensions of Biodiversity Program (DEB-1542509). Sequencing was conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, was supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Supplementary Materials

Figure S1 Pairwise Scaled Identity by State (IBS) for the (a) complete association population (N=112), (b) the complete F_2 full sib Family (N=497), and (c) the association population with clones removed (N=75). The IBS cutoff used for identifying clonal pairs was 0.9.

Figure S2 Frequency of mapped markers with and without segregation distortion in family 317 for males and females. A. Markers in female-backcross configuration. B. Markers in male-backcross configuration. Notice the lack of undistorted (normal) markers on chromosome 19 in female backcross configuration.

Figure S3 Quantile–Quantile (Q–Q) plots of observed and expected P-values for the GWAS for sex. Red line indicates X = Y.

Figure S4 Stacked histogram of average observed heterozygosity for males, females, and hermaphrodites for sex-associated loci in the *S. purpurea* association population.

Figure S5 Distribution of differences in null allele frequency between females and males in the association population. Extreme values are shaded in red.

Figure S6 Proportion of reference sequence gaps ("assembly Ns") in regions that showed no coverage in the female (a) or male (b) reference-based alignments. The male had 0 coverage primarily in regions with minimal reference gaps, suggesting that these are regions that are present in the female sequence and absent in the male.

Figure S7 Box plot showing that the proportion of repeat elements is elevated in the SDR.

Figure S8 Dot plot derived from aligning the *S. suchowensis* SDR (primarily located on scaffold64) to *S. purpurea* chromosome 15 using lastz.

Figure S9 Alignment of Kinesin genes from the SDR of *S. purpurea* and their closest ortholog in *P. trichocarpa.* SapurV1A.1267s0010 is artificially truncated due to an assembly gap overlapping with the gene. Conserved domains are highlighted and labeled. Tandem duplicate pairs are 1.) SapurV1A.0719s0080 and SapurV1A.0719s0090; and 2.) SapurV1A.1267s0010 and SapurV1A.1267s0020.

Table S1 Significant markers (LOD>3.5) from QTL mapping of sex. The table includes linkage group (LG), map positions (in centimorgans), map type (female backcross, F, male backcross, M, and intercross, IC), the physical scaffold from the genome assembly, the physical position of the marker in the genome assembly, and the frequency of different genotype configurations in the progeny.

Table S2 Number of unfiltered GBS markers produced by the Tassel pipeline for the F_2 family 317. Markers/100kb is the average number of markers per 100 kb interval. F:M Backcross is the ratio of markers in a Female Backcross configuration (heterozygous in the female parent, homozygous in the male parent) to markers in the Male Backcross configuration (homozygous in female parent, heterozygous in male parent).

Table S3 Results of GWAS for sex. The table includes all significant markers ($p<1x10^{-7}$).

Table S4 Best matches for secondary *S. purpurea* SDRs to the *S. purpurea* and *P. trichocarpa* genomes. "Secondary Blast Hit" is the best blastn hit to the *S. purpurea* genome, after excluding self hits.

Table S5 Markers showing a female-specific genotype configuration (one allele observed in females, none in males). These are presumably derived from W segments included in the genome assembly.

Table S6 Scaffolds with >30% female-specific sequence. "Proportion W" is a calculation based on the proportion of the scaffold, after excluding gaps, that is present in the female sequence but absent in the male sequence (Female-Specific).

Table S7 Repeat composition of the *S. purpurea* chromosomes.

Table S8 Predicted genes found within the SDR of *S. purpurea*. "W Overlap" and "W proportion" represent the intersection of the location of the gene with female-specific genome segments. Omega values are the ratio of nonsynonymous (dN) to synonymous (dS) substitutions

687 between the *S. purpurea* and *P. trichocarpa* orthologs. Multiple values are provided in cases
688 with multiple *Populus* orthologs, presumably due to lineage-specific expansion.
689

690 ## Literature Cited
691

692 Ainsworth, C., 2000 Boys and girls come out to play: The molecular biology of dioecious plants.
693 Ann. Bot. 86: 211–221.

694 Alliende, M. C., and J. L. Harper, 1989 Demographic studies of a dioecious tree. I. Colonization,
695 sex and age structure of a population of *Salix cinerea*. J. Ecol. 77: 1029–1047.

696 Alstrom-Rapaport, C., M. Lascoux, Y. C. Wang, G. Roberts, and G. A. Tuskan, 1998
697 Identification of a RAPD marker linked to sex determination in the basket willow (*Salix*
698 *viminalis* L.). J. Hered. 89: 44–49.

699 Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe, 2016 Harnessing
700 the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet 17: 81–92.

701 Arends, D., P. Prins, R. C. Jansen, and K. W. Broman, 2010 R/qtl: High-throughput multiple
702 QTL mapping. Bioinformatics 26: 2990–2992.

703 Argus, G. W., 1997 Infrageneric classification of *Salix* (Salicaceae) in the new world. Syst. Bot.
704 Monogr. 52: 1–121.

705 Ashman, T.-L., 2006 The evolution of separate sexes: a focus on the ecological context, pp. 370
706 in *Ecology and evolution of flowers*, edited by L. D. Harder and S. C. H. Barrett. Oxford
707 University Press, Oxford.

708 Ashman, T.-L., A. Kwok, and B. C. Husband, 2013 Revisiting the Dioecy-Polyploidy
709 Association: Alternate Pathways and Research Opportunities. Cytogenet. Genome Res. 140:
710 241–255.

711 Bachtrog, D., J. Mank, C. L. Peichel, M. Kirkpatrick, S. P. Otto *et al.*, 2014 Sex Determination:
712 Why So Many Ways of Doing It? PLoS Biol. 12: e1001899.

713 Barrett, S. C. H., and J. Hough, 2013 Sexual dimorphism in flowering plants. J. Exp. Bot. 64:
714 67–82.

715 Bergero, R., and D. Charlesworth, 2009 The evolution of restricted recombination in sex
716 chromosomes. Trends Ecol. Evol. 24: 94–102.

717 Bergero, R., A. Forrest, E. Kamau, and D. Charlesworth, 2007 Evolutionary strata on the X
718 chromosomes of the dioecious plant *Silene latifolia*: Evidence from new sex-linked genes.
719 Genetics 175: 1945–1954.

720 Boucher, L. D., S. R. Manchester, and W. S. Judd, 2003 An extinct genus of Salicaceae based on
721 twigs with attached flowers, fruits, and foliage from the Eocene Green River Formation of
722 Utah and Colorado, USA. Am. J. Bot. 90: 1389–1399.

723 Burby, P. E., and L. A. Simmons, 2017 MutS2 promotes homologous recombination in *Bacillus*
724 *subtilis*. J. Bacteriol. 199: e00682-16.

725 Cai, G., and M. Cresti, 2009 Organelle motility in the pollen tube: a tale of 20 years. J. Exp. Bot.
726 60: 495–508.

727 Carlson, C. H., Y. Choi, A. P. Chan, M. J. Serapiglia, C. D. Town *et al.*, 2017 Dominance and
728 Sexual Dimorphism Pervade the *Salix purpurea* L. Transcriptome. Genome Biol. Evol. 9:
729 2377–2394.

730 Charlesworth, D., 2006 Evolution of Plant Breeding Systems. Curr. Biol. 16: 726–735.

731 Charlesworth, D., 2015 Plant contributions to our understanding of sex chromosome evolution.
732 New Phytol. 208: 52–65.

733 Charlesworth, D., 2016 Plant Sex Chromosomes. Annu. Rev. Plant Biol. 67: 397–420.

734 Charlesworth, D., and B. Charlesworth, 1978 Population genetics of partial male-sterility and the
735 evolution of monoecy and dioecy. Heredity 41: 137–153.

736 Charnov, E. L., 1982 The theory of sex allocation. Monogr. Popul. Biol. 18: 1–355.

737 Chen, Y., T. Wang, L. Fang, X. Li, and T. Yin, 2016 Confirmation of single-locus sex
738 determination and female heterogamety in willow based on linkage analysis. PLoS One 11:
739 e0147671.

740 DePristo, M. a, E. Banks, R. Poplin, K. V Garimella, J. R. Maguire *et al.*, 2011 A framework for
741 variation discovery and genotyping using next-generation DNA sequencing data. Nat.
742 Genet. 43: 491–8.

743 Dickmann, D. I., and J. Kuzovkina, 2014 Poplars and willows of the world, with emphasis on
744 silviculturally important species., pp. 8–91 in *Poplars and willows: trees for society and the*
745 *environment*, CABI, Wallingford.

746 van Doorn, G. S., and M. Kirkpatrick, 2010 Transitions between male and female heterogamety
747 caused by sex-antagonistic selection. Genetics 186: 629–645.

748 van Doorn, G. S., and M. Kirkpatrick, 2007 Turnover of sex chromosomes induced by sexual
749 conflict. Nature 449: 909–912.

750 Elshire, R. J., J. C. Glaubitz, Q. Sun, J. a. Poland, K. Kawamoto *et al.*, 2011 A robust, simple
751 genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379.

752 Falda, M., S. Toppo, A. Pescarolo, E. Lavezzo, B. Di Camillo *et al.*, 2012 Argot2: a large scale
753 function prediction tool relying on semantic similarity of weighted Gene Ontology terms.
754 BMC Bioinformatics 13: S14.

755 Forsthoefel, N., K. Klag, B. Simeles, R. Reiter, L. Brougham *et al.*, 2013 The *Arabidopsis* Plant
756 Intracellular Ras-group LRR (PIRL) Family and the Value of Reverse Genetic Analysis for
757 Identifying Genes that Function in Gametophyte Development. Plants 2: 507–520.

758 Fukui, K., H. Kosaka, S. Kuramitsu, and R. Masui, 2007 Nuclease activity of the MutS
759 homologue MutS2 from *Thermus thermophilus* is confined to the Smr domain. Nucleic
760 Acids Res. 35: 850–860.

761 Fukui, K., and S. Kuramitsu, 2011 Structure and Function of the Small MutS-Related Domain.
762 Mol. Biol. Int. 2011: 1–9.

763 Füssel, U., S. Dötterl, A. Jürgens, and G. Aas, 2007 Inter- and Intraspecific Variation in Floral
764 Scent in the Genus *Salix* and its Implication for Pollination. J. Chem. Ecol. 33: 749–765.

765 Gaudet, M., V. Jorge, I. Paolucci, I. Beritognolo, G. S. Mugnozza *et al.*, 2008 Genetic linkage
766 maps of *Populus nigra* L. including AFLPs, SSRs, SNPs, and sex trait. Tree Genet.
767 Genomes 4: 25–36.

768 Geraldes, A., C. A. Hefer, A. Capron, N. Kolosova, F. Martinez-Nuñez *et al.*, 2015 Recent Y
769 chromosome divergence despite ancient origin of dioecy in poplars (*Populus*). Mol. Ecol.
770 24: 3243–3256.

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire *et al.*, 2014 TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS One 9: e0090346.

Glick, L., N. Sabath, T. L. Ashman, E. Goldberg, and I. Mayrose, 2016 Polyploidy and sexual system in angiosperms: Is there an association? Am. J. Bot. 103: 1223–1235.

Gnerre, S., I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton *et al.*, 2011 High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U. S. A. 108: 1513–1518.

Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes *et al.*, 2012 Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40: D1178–D1186.

Hou, J., N. Ye, Z. Dong, M. Lu, L. Li *et al.*, 2016 Major chromosomal rearrangements distinguish willow and poplar after the ancestral "Salicoid" genome duplication. Genome Biol. Evol. 8: 1868–1875.

Hou, J., N. Ye, D. Zhang, Y. Chen, L. Fang *et al.*, 2015 Different autosomes evolved into sex chromosomes in the sister genera of *Salix* and *Populus*. Sci. Rep. 5: e9076.

Hsieh, K., and A. H. C. Huang, 2004 Endoplasmic reticulum, oleosins, and oils in seeds and tapetum cells. Plant Physiol. 136: 3427–3434.

Kang, H. M., J. H. Sul, S. K. Service, N. a Zaitlen, S.-Y. Kong *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42: 348–354.

Karp, A., S. J. Hanley, S. O. Trybush, W. Macalpine, M. Pei *et al.*, 2011 Genetic Improvement of Willow for Bioenergy and Biofuels. J. Integr. Plant Biol. 53: 151–165.

Karrenberg, S., J. Kollmann, and P. J. Edwards, 2002 Pollen vectors and inflorescence morphology in four species of *Salix*. Plant Syst. Evol. 235: 181–188.

Kersten, B., B. Pakull, K. Groppe, J. Lueneburg, and M. Fladung, 2014 The sex-linked region in *Populus tremuloides* Turesson 141 corresponds to a pericentromeric region of about two million base pairs on *P. trichocarpa* chromosome 19. Plant Biol. 16: 411–418.

Kim, H. U., K. Hsieh, C. Ratnayake, and A. H. C. Huang, 2002 A novel group of oleosins is present inside the pollen of *Arabidopsis*. J. Biol. Chem. 277: 22677–22684.

Kunkel, T. a, and D. a Erie, 2005 DNA mismatch repair. Annu. Rev. Biochem. 74: 681–710.

Lian, S., T. Liu, K. Gong, X. Chen, and G. Zheng, 2016 A Complete and Accurate Short Sequence Alignment Algorithm for Repeats. J. Biosci. Med. 4: 144–151.

Lin, H., L. Niu, N. a McHale, M. Ohme-Takagi, K. S. Mysore *et al.*, 2013 Evolutionarily conserved repressive activity of WOX proteins mediates leaf blade outgrowth and floral organ development in plants. Proc. Natl. Acad. Sci. U. S. A. 110: 366–371.

Lloyd, D. G., 1979 Evolution towards dioecy in heterostylous populations. Plant Syst. Evol. 131: 71–80.

Mank, J. E., 2009 Sex chromosomes and the evolution of sexual dimorphism: lessons from the genome. Am. Nat. 173: 141–150.

Melters, D. P., K. R. Bradnam, H. a Young, N. Telis, M. R. May *et al.*, 2013 Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 14: R10.

Miller, J. R., S. Koren, and G. Sutton, 2010 Assembly algorithms for next-generation sequencing data. Genomics 95: 315–27.

Ming, R., A. Bendahmane, and S. S. Renner, 2011 Sex chromosomes in land plants. Annu. Rev. Plant Biol. 62: 485–514.

Ming, R., and P. H. Moore, 2007 Genomics of sex chromosomes. Curr. Opin. Plant Biol. 10: 123–130.

Mock, K. E., C. M. Callahan, M. N. Islam-Faridi, J. D. Shaw, H. S. Rai *et al.*, 2012 Widespread triploidy in western North American aspen (*Populus tremuloides*). PLoS One 7: e48406.

Moore, E. C., and R. B. Roberts, 2013 Polygenic sex determination. Curr. Biol. 23: R510-2.

Nicolas, M., G. Marais, V. Hykelova, B. Janousek, V. Laporte *et al.*, 2005 A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. PLoS Biol. 3: e4.

Nishihama, R., T. Soyano, M. Ishikawa, S. Araki, H. Tanaka *et al.*, 2002 Expansion of the cell plate in plant cytokinesis requires a kinesin-like protein/MAPKKK complex. Cell 109: 87–99.

827 Olson, M. S., J. L. Hamrick, and R. C. Moore, 2017 Breeding systems, mating systems, and
828 gender determination in angiosperm trees, in *Comparative and Evolutionary Genomics of*
829 *Angiosperm Trees*, edited by A. Groover and Q. C. B. Cronk. Springer International
830 Publishing, Switzerland.

831 Otto, S. P., J. R. Pannell, C. L. Peichel, T.-L. Ashman, D. Charlesworth *et al.*, 2011 About PAR:
832 the distinct evolutionary dynamics of the pseudoautosomal region. Trends Genet. 27: 358–
833 367.

834 Pakull, B., K. Groppe, M. Meyer, T. Markussen, and M. Fladung, 2009 Genetic linkage mapping
835 in aspen (*Populus tremula* L. and *Populus tremuloides* Michx.). Tree Genet. Genomes 5:
836 505–515.

837 Pakull, B., B. Kersten, J. Lüneburg, and M. Fladung, 2014 A simple PCR-based marker to
838 determine sex in aspen. Plant Biol. 17: 256–261.

839 Pandey, R. S., and R. K. Azad, 2016 Deciphering evolutionary strata on plant sex chromosomes
840 and fungal mating-type chromosomes through compositional segmentation. Plant Mol. Biol.
841 90: 359–373.

842 Paolucci, I., M. Gaudet, V. Jorge, I. Beritognolo, S. Terzoli *et al.*, 2010 Genetic linkage maps of
843 *Populus alba* L. and comparative mapping analysis of sex determination across Populus
844 species. Tree Genet. Genomes 6: 863–875.

845 Peto, F. H., 1938 Cytology of poplar species and natural hybrids. Can. J. Res. 16: 446–455.

846 Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. a Shadick *et al.*, 2006 Principal
847 components analysis corrects for stratification in genome-wide association studies. Nat.
848 Genet. 38: 904–909.

849 Pucholt, P., H. R. Hallingbäck, and S. Berlin, 2017 Allelic incompatibility can explain female
850 biased sex ratios in dioecious plants. BMC Genomics 18: 251.

851 Pucholt, P., A.-C. Rönnberg-Wästljung, and S. Berlin, 2015 Single locus sex determination and
852 female heterogamety in the basket willow (*Salix viminalis* L.). Heredity 114: 575–583.

853 Pucholt, P., A. E. Wright, L. L. Conze, J. E. Mank, and S. Berlin, 2017 Recent Sex Chromosome
854 Divergence despite Ancient Dioecy in the Willow, *Salix viminalis*. Mol. Biol. Evol. 22:

522–525.

Renner, S. S., 2014 The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. Am. J. Bot. 101: 1588–1596.

Rice, W. W. R., 1984 Sex chromosomes and the evolution of sexual dimorphism. Evolution 38: 1416–1424.

Serapiglia, M. J., F. E. Gouker, J. F. Hart, F. Unda, S. D. Mansfield *et al.*, 2015 Ploidy Level Affects Important Biomass Traits of Novel Shrub Willow (*Salix*) Hybrids. BioEnergy Res. 8: 259–269.

Slavov, G. T., and P. Zhelev, 2010 Salient Biological Features, Systematics, and Genetic Variation of *Populus*, pp. 15–38 in *Genetics and Genomics of* Populus, edited by S. Jansson, R. P. Bhalerao, and A. Groover. Springer New York, New York, NY.

Sterck, L., S. Rombauts, S. Jansson, F. Sterky, P. Rouzé *et al.*, 2005 EST data suggest that poplar is an ancient polyploid. New Phytol. 167: 165–170.

Temmel, N. A., H. S. Rai, and Q. C. B. Cronk, 2007 Sequence characterization of the putatively sex-linked Ssu72 -like locus in willow and its homologue in poplar. Can. J. Bot. 85: 1092–1097.

Treangen, T. J., and S. L. Salzberg, 2013 Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 13: 36–46.

Tuskan, G. A., S. DiFazio, P. Faivre-Rampant, M. Gaudet, A. Harfouche *et al.*, 2012 The obscure events contributing to the evolution of an incipient sex chromosome in *Populus*: a retrospective working hypothesis. Tree Genet. Genomes 8: 559–571.

Tuskan, G. A., S. DiFazio, S. Jansson, J. Bohlmann, I. Grigoriev *et al.*, 2006 The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313: 1596–1604.

Ueno, N., Y. Suyama, and K. Seiwa, 2007 What makes the sex ratio female-biased in the dioecious tree *Salix sachalinensis*? J. Ecol. 95: 951–959.

Vyskot, B., and R. Hobza, 2015 The genomics of plant sex chromosomes. Plant Sci. 236: 126–135.

Wang, J., J. Na, Q. Yu, A. R. Gschwend, J. Han *et al.*, 2012 Sequencing papaya X and Y h

chromosomes reveals molecular basis of incipient sex chromosome evolution. Proc. Natl. Acad. Sci. 109: 13710–13715.

Yang, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24: 1586–1591.

Yang, C., G. Vizcay-Barrena, K. Conner, and Z. a Wilson, 2007 MALE STERILITY1 is required for tapetal development and pollen wall biosynthesis. Plant Cell 19: 3530–3548.

Yin, T., S. P. DiFazio, L. E. Gunter, X. Zhang, M. M. Sewell *et al.*, 2008 Genome structure and emerging evidence of an incipient sex chromosome in *Populus*. Genome Res. 18: 422–430.

Zhou, W., Q. Lu, Q. Li, L. Wang, S. Ding *et al.*, 2017 PPR-SMR protein SOT1 has RNA endonuclease activity. Proc. Natl. Acad. Sci. U. S. A. 114: E1554–E1563.